*Published online 24 November 2012* *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D861–D864***
*doi:10.1093/nar/gks946*

# The Zebrafish Insertion Collection (ZInC): a web based, searchable collection of zebrafish mutations generated by DNA insertion

**Gaurav K. Varshney[1], Haigen Huang[2], Suiyuan Zhang[1], Jing Lu[2], Derek E. Gildea[1], Zhongan Yang[2], Tyra G. Wolfsberg[1], Shuo Lin[2,]* and Shawn M. Burgess[1,]***

[1]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA and [2]Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles (UCLA), Los Angeles, USA

Received August 15, 2012; Revised September 14, 2012; Accepted September 19, 2012

## ABSTRACT

**ZInC (Zebrafish Insertional Collection, http://research.nhgri.nih.gov/ZInC/) is a web-searchable interface of insertional mutants in zebrafish. Over the last two decades, the zebrafish has become a popular model organism for studying vertebrate development as well as for modeling human diseases. To facilitate such studies, we are generating a genome-wide knockout resource that targets every zebrafish protein-coding gene. All mutant fish are freely available to the scientific community through the Zebrafish International Resource Center (ZIRC). To assist researchers in finding mutant and insertion information, we developed a comprehensive database with a web front-end, the ZInC. It can be queried using multiple types of input such as ZFIN (Zebrafish Information Network) IDs, UniGene accession numbers and gene symbols from zebrafish, human and mouse. In the future, ZInC may include data from other insertional mutation projects as well. ZInC cross-references all integration data with the ZFIN (http://zfin.org/).**

## INTRODUCTION

Functional genomic studies in different model organisms including vertebrates such as mice, frogs and fish have contributed immensely to our understanding of various human diseases. Furthermore, it is now possible to systematically create gene mutations and then study the associated phenotypes (i.e., 'reverse' genetics). Zebrafish has gained momentum in the past two decades as a genetically tractable model organism in which to study vertebrate development because of its transparent embryos, small size, *in vitro* development, high fecundity and inexpensive maintenance (1). Their embryonic development is fast, with most of the critical development occurring in the first 5 days. Traditionally in zebrafish, forward genetic screens are used, and two independent, large-scale mutagenesis screens have been carried out using ENU as the mutagen (2,3). ENU is very efficient, but it introduces single base-pair change into DNA, and the identification of the target gene is typically done by a rather laborious and time-consuming method of positional cloning (4,5). To circumvent the positional cloning step, insertional mutagenesis strategies have been developed in different model organisms. Insertional elements such as transposons and retroviruses have become indispensible tools in manipulating genomes for various applications, not only insertional mutagenesis, but also transgenesis and gene therapy (6,7). Insertional mutagenesis using retroviral vectors has effectively been used to disrupt gene functions in vertebrates (8,9). We are using a high-throughput retroviral mediated mutagenesis followed by mapping using next-generation sequencing methods (10) to generate a knockout library of zebrafish. In order to make this mutagenic resource readily available to the research community, we have developed ZInC, the Zebrafish Integration Collection, an integrated database and web front-end to display mutagenic insertions in a simple and interactive way.

## INSERTION DATA SOURCE

The experimental data in ZInC are derived from our ongoing retroviral mediated insertional mutagenesis project (Varshney and Lu et al., unpublished). A graphical representation of the pipeline is included in the ZInC website.

*To whom correspondence should be addressed. Tel: +301 594 8224; Fax: +301 496 0474; Email: burgess@mail.nih.gov
Correspondence may also be addressed to: Shuo Lin. Email: shuolin@ucla.edu

## INSERTION DATA PROCESSING

We developed a robust mapping analysis pipeline to generate insertion data in the zebrafish genome. Raw data were processed from ELAND or BAM output files, generated from paired-end sequencing using Illumina platform, and sequence reads were extracted. To map insertion sites, retroviral vector and linker sequences were trimmed, then the trimmed reads were mapped to zebrafish genome (Ensembl Zv9.0 assembly) using the short read aligner Bowtie (12). Since LM-PCR was performed to isolate specific integration events, it was possible that the same integration site was sequenced multiple times. Therefore, sequence reads that mapped to the same chromosome, having the same genomic alignment start position, and mapping to the same DNA strand defined a single integration event. The mapped integration sites were then compared to the genomic locations of annotated genes (Ensembl e65) to determine which integration sites are associated with genes, that is, which integration sites are within exons or introns. Insertion events are flagged as being 'Predicted Mutagenic' only if they fall within any exon or the first intron.

## DATABASE DESIGN AND IMPLEMENTATION

The ZInC website interface uses a Common Gateway Interface (CGI) constructed in the Perl programming language that interacts with a relational database hosted in Oracle 11g. The web interface was developed by using HTML, Perl, Java Script and Template Toolkit. The connectivity between the CGI and the Oracle 11g Relational Database Management Software was implemented using Perl's Database Interface (DBI) and the Oracle database driver for the DBI module (DBD::Oracle). The ZInC database consists of several tables that hold the data content, including the zebrafish gene name and gene symbols, human and mouse gene orthologs, and zebrafish KEGG (Kyoto Encyclopedia of Genes and Genomes) information (13). The tables are populated by downloading data from DAVID (http://david.abcc.ncifcrf.gov/; KEGG only) (14,15) and Zebrafish Information Network (ZFIN) (http://zfin.org/; all other annotations) (16); these data will be refreshed quarterly, or as available. The list of insertions is also updated as new experimental and bioinformatics analyses become available. A suite of Perl scripts was developed to add new data and annotations into ZInC as well as to ensure data integrity.

## DATABASE NAVIGATION

The navigation sidebar on the left side of each page provides links to different sections of the web resource. As shown in Figure 1, the 'Search' link is the main one, a search page that provides access to the integration sites in the database. To facilitate the search, we provide a simple interface that accepts multiple input IDs such as Ensembl (e.g. ENSDARG00000010070), Genbank (e.g. BC081408), RefSeq (e.g. NM_001004678), UniGene (e.g. Dr.134464) or ZFIN (e.g. ZDB-GENE-040912-127) identifiers. The user can enter either a single ID or a list of IDs separated by commas, spaces or carriage returns. Users can also search by gene symbol (e.g. smo) or gene name (e.g. smoothened homolog) from zebrafish, human or mouse. A list of gene names or gene symbols can also be used as search input; again, the list of identifiers can be separated by comma, space or carriage returns. Researchers studying biochemical pathways can search by KEGG pathway (13) terms (e.g. Glycosphingolipid biosynthesis) to find insertions in genes in a specific pathway.

The results of a search are shown in Figure 2. For each gene hit by the query term, the ZFIN ID, zebrafish gene symbol and zebrafish gene name are returned, regardless of whether that gene is disrupted by an integration. The presence of an integration is indicated by the link 'View integration' in the Integration column; whether the insertion is predicted to be mutagenic (i.e. it lands in an exon or first intron) is marked in the final column. The results of the 'View integration' link are shown in Figure 3. In brief, this page shows the integration position, allele number, Ensembl gene ID and a link to order fish through the Zebrafish International Resource Center (ZIRC) when available. The insertion position is linked to the UCSC Genome Browser so that users can see the genomic context of each integration site. Each allele number is linked to the corresponding ZFIN allele page.

## CONCLUSIONS

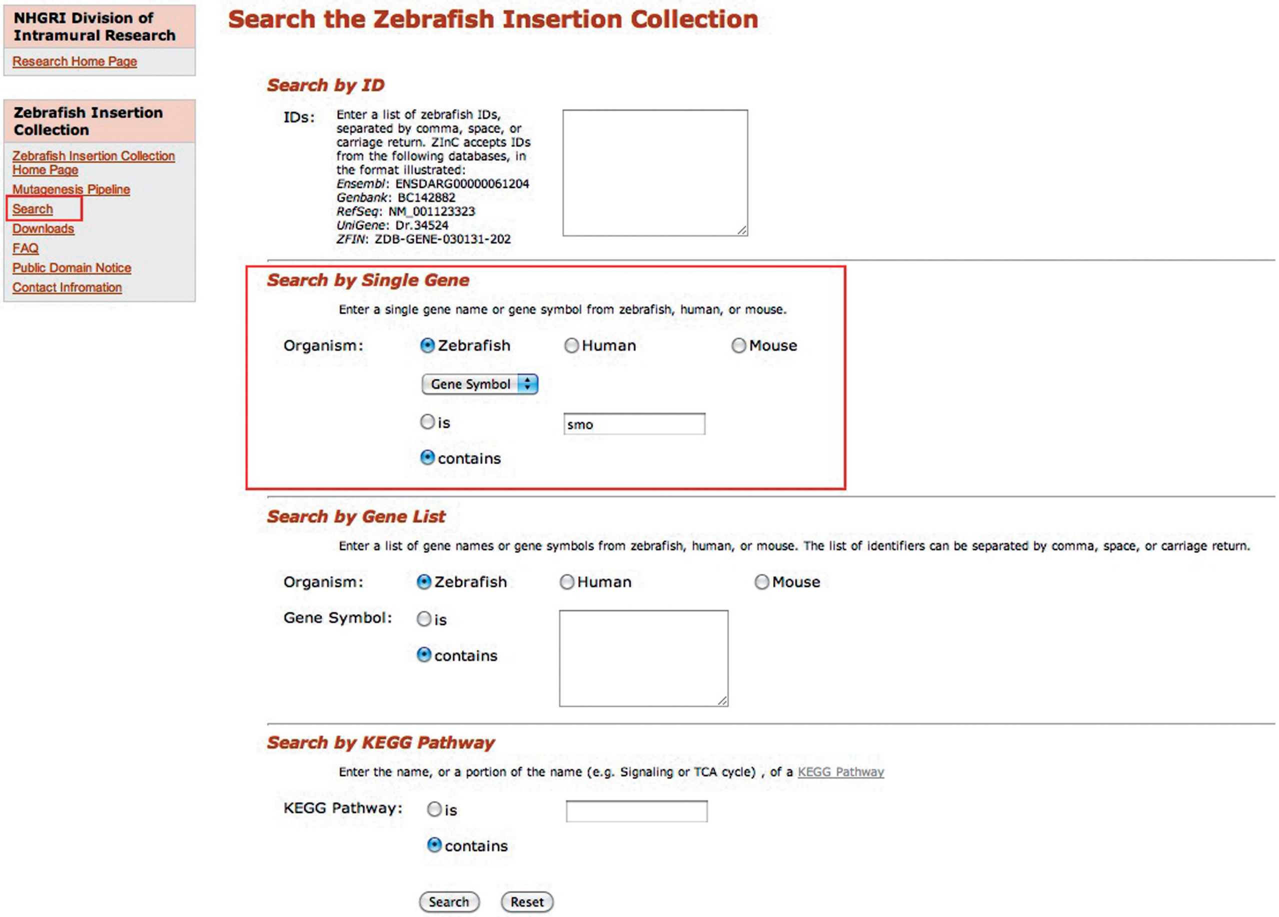
ZInC is a part of an ongoing project where we aim to knock out every protein-coding gene in the zebrafish genome. We will update the database at least quarterly with newly identified integration sites. Other groups are also attempting to knock out genes using different insertional elements, such as the *Tol2* and *Ac*/*Ds* transposons (17,18). An effort is being made to integrate these and other similar data into ZInC, allowing it to serve as a central repository for all integration sites in zebrafish.

## ACCESSIBILITY

ZInC can be accessed at http://research.nhgri.nih.gov/zinc. Comprehensive lists of all insertion sites, as well as protocols and methods required for the genotyping can be downloaded from http://research.nhgri.nih.gov/zinc/?mode=downloads. All the data cross-references with gene and mutant data in the ZFIN: http://zfin.org.

## ACCESSION NUMBERS

As of August 2012, 13 316 insertion sequences have been submitted to the National Center for Biotechnology Information (NCBI) Genome Survey Sequence (GSS) database. The accession numbers are JS426363-JS426454, JS495495-JS496658, JS578512-JS583384, JS672208-JS672893, JS784708-JS785225 and JS876947-JS886733, and the BioSample ID is GSS: LIBGSS_038780. The full list of integrations is available from the Downloads page at: http://research.nhgri.nih.gov/zinc/?mode=downloads.



**NHGRI Division of Intramural Research**

Research Home Page

**Zebrafish Insertion Collection**

Zebrafish Insertion Collection Home Page

Mutagenesis Pipeline

Search

Downloads

FAQ

Public Domain Notice

Contact Infromation

**Search the Zebrafish Insertion Collection**

***Search by ID***

IDs: Enter a list of zebrafish IDs, separated by comma, space, or carriage return. ZInC accepts IDs from the following databases, in the format illustrated:
*Ensembl*: ENSDARG00000061204
*Genbank*: BC142882
*RefSeq*: NM_001123323
*UniGene*: Dr.34524
*ZFIN*: ZDB-GENE-030131-202

***Search by Single Gene***

Enter a single gene name or gene symbol from zebrafish, human, or mouse.

Organism: ● Zebrafish ○ Human ○ Mouse

Gene Symbol

○ is smo

● contains

***Search by Gene List***

Enter a list of gene names or gene symbols from zebrafish, human, or mouse. The list of identifiers can be separated by comma, space, or carriage return.

Organism: ● Zebrafish ○ Human ○ Mouse

Gene Symbol: ○ is

● contains

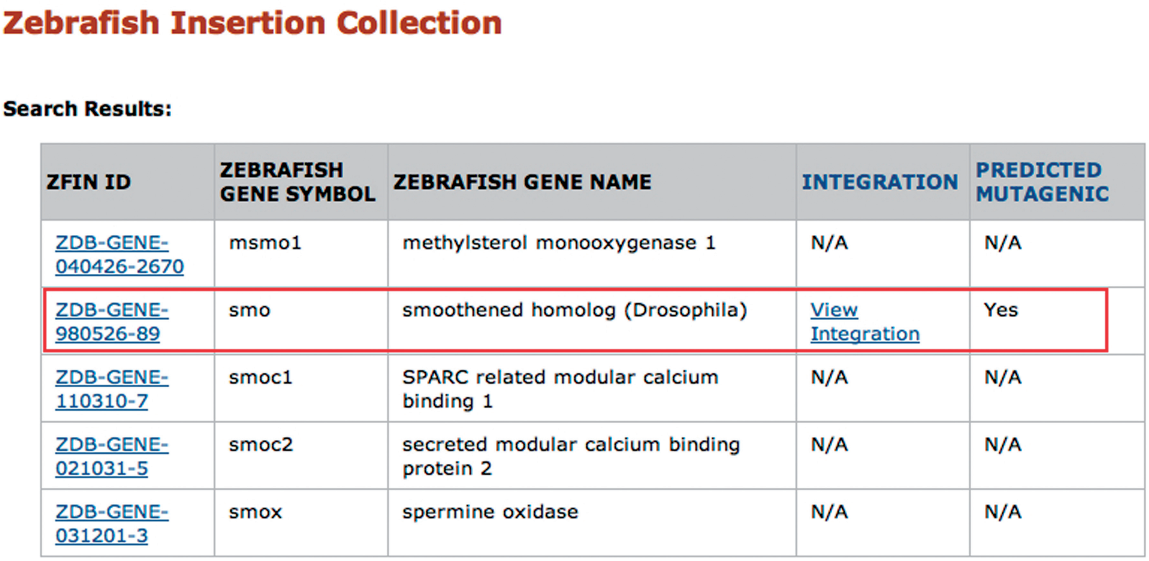
***Search by KEGG Pathway***

Enter the name, or a portion of the name (e.g. Signaling or TCA cycle) , of a KEGG Pathway

KEGG Pathway: ○ is

● contains

**Figure 1.** ZInC search interface. Insertion sites within genes have been mapped to a variety of common identifiers. Users can query ZInC with accession numbers from a number of sources, including Ensembl, GenBank, RefSeq and ZFIN. Users can also query by human, mouse or zebrafish gene symbols and names, either individual entries or longer lists. Queries can also be performed on KEGG biological pathways. All searches allow for an exact match (is) or a query with wildcards (contains). In this instance, we searched for a mutant in the gene *smoothened* using the zebrafish symbol 'smo' and by choosing the 'contains' radio button, the search will return any gene symbol that has the text string 'smo' in it.

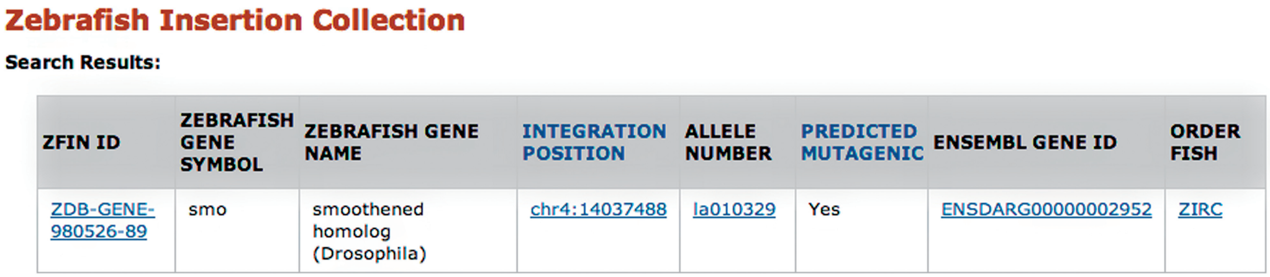


**Zebrafish Insertion Collection**

**Search Results:**

| ZFIN ID | ZEBRAFISH GENE SYMBOL | ZEBRAFISH GENE NAME | INTEGRATION | PREDICTED MUTAGENIC |
|---|---|---|---|---|
| ZDB-GENE-040426-2670 | msmo1 | methylsterol monooxygenase 1 | N/A | N/A |
| ZDB-GENE-980526-89 | smo | smoothened homolog (Drosophila) | View Integration | Yes |
| ZDB-GENE-110310-7 | smoc1 | SPARC related modular calcium binding 1 | N/A | N/A |
| ZDB-GENE-021031-5 | smoc2 | secreted modular calcium binding protein 2 | N/A | N/A |
| ZDB-GENE-031201-3 | smox | spermine oxidase | N/A | N/A |

**Figure 2.** ZInC search results. In Figure 1, the zebrafish gene symbol 'smo' was entered in the Search by Single Gene box. Since the 'contains' radio button was selected, all zebrafish gene symbols in ZFIN containing the text string '*smo*' are returned, regardless of whether an integration in the gene is available. IDs in the ZFIN ID column link to ZFIN entries for specific genes. For those genes with an integration, more detailed information is available through the 'View Integration' link (Figure 3).



Zebrafish Insertion Collection

Search Results:

| ZFIN ID | ZEBRAFISH GENE SYMBOL | ZEBRAFISH GENE NAME | INTEGRATION POSITION | ALLELE NUMBER | PREDICTED MUTAGENIC | ENSEMBL GENE ID | ORDER FISH |
|---|---|---|---|---|---|---|---|
| ZDB-GENE-980526-89 | smo | smoothened homolog (Drosophila) | chr4:14037488 | la010329 | Yes | ENSDARG00000002952 | ZIRC |

**Figure 3.** ZInC integration site details. Clicking on the 'View Integration' link on a search results page (Figure 2) results in a detailed view of the integration site. The 'Integration Position' column links to the UCSC Genome Browser zoomed in to a 2 Kb window around the integration site, the 'Ensembl Gene ID' column links to the Ensembl gene page, and the 'ORDER FISH' column links directly to the ZIRC to purchase the desired mutant fish (if available).

## ACKNOWLEDGEMENTS

We would like to thank Colin Huck for outstanding animal care, and Robert Blakesley, Alice Young and the NIH Intramural Sequencing Center (NISC) for providing sequence data.

## FUNDING

Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (to S.B. and T.W.) and R01 grant NIH DK084349 (to S.L.). Funding for open access charge: Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (to S.B. and T.W.) and R01 grant NIH DK084349 (to S.L.).

*Conflict of interest statement*. None declared.

## REFERENCES

1. Kimmel,C.B. (1989) Genetics and early development of zebrafish. *Trends Genet*., **5**, 283–288.
2. Haffter,P., Granato,M., Brand,M., Mullins,M.C., Hammerschmidt,M., Kane,D.A., Odenthal,J., van Eeden,F.J., Jiang,Y.J., Heisenberg,C.P. *et al.* (1996) The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio. *Development*, **123**, 1–36.
3. Driever,W., Solnica-Krezel,L., Schier,A.F., Neuhauss,S.C., Malicki,J., Stemple,D.L., Stainier,D.Y., Zwartkruis,F., Abdelilah,S., Rangini,Z. *et al.* (1996) A genetic screen for mutations affecting embryogenesis in zebrafish. *Development*, **123**, 37–46.
4. Solnica-Krezel,L., Schier,A.F. and Driever,W. (1994) Efficient recovery of ENU-induced mutations from the zebrafish germline. *Genetics*, **136**, 1401–1420.
5. Talbot,W.S. and Schier,A.F. (1999) Positional cloning of mutated zebrafish genes. *Methods Cell Biol*., **60**, 259–286.
6. Jao,L.E., Maddison,L., Chen,W. and Burgess,S.M. (2008) Using retroviruses as a mutagenesis tool to explore the zebrafish genome. *Brief. Funct. Genomic. Proteomic*., **7**, 427–443.
7. Sivasubbu,S., Balciunas,D., Amsterdam,A. and Ekker,S.C. (2007) Insertional mutagenesis strategies in zebrafish. *Genome Biol*., **8(Suppl. 1)**, S9.
8. Gaiano,N., Amsterdam,A., Kawakami,K., Allende,M., Becker,T. and Hopkins,N. (1996) Insertional mutagenesis and rapid cloning of essential genes in zebrafish. *Nature*, **383**, 829–832.
9. Amsterdam,A., Burgess,S., Golling,G., Chen,W., Sun,Z., Townsend,K., Farrington,S., Haldi,M. and Hopkins,N. (1999) A large-scale insertional mutagenesis screen in zebrafish. *Genes Dev*., **13**, 2713–2724.
10. Wang,D., Jao,L.E., Zheng,N., Dolan,K., Ivey,J., Zonies,S., Wu,X., Wu,K., Yang,H., Meng,Q. *et al.* (2007) Efficient genome-wide mutagenesis of zebrafish genes by retroviral insertions. *Proc. Natl Acad. Sci. USA*, **104**, 12428–12433.
11. Jao,L.E. and Burgess,S.M. (2009) Production of pseudotyped retrovirus and the generation of proviral transgenic zebrafish. *Methods Mol. Biol*., **546**, 13–30.
12. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*., **10**, R25.
13. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
14. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols*, **4**, 44–57.
15. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*., **37**, 1–13.
16. Bradford,Y., Conlin,T., Dunn,N., Fashena,D., Frazer,K., Howe,D.G., Knight,J., Mani,P., Martin,R., Moxon,S.A. *et al.* (2011) ZFIN: enhancements and updates to the Zebrafish Model Organism Database. *Nucleic Acids Res*., **39**, D822–D829.
17. Emelyanov,A., Gao,Y., Naqvi,N.I. and Parinov,S. (2006) Trans-kingdom transposition of the maize dissociation element. *Genetics*, **174**, 1095–1104.
18. Kawakami,K., Takeda,H., Kawakami,N., Kobayashi,M., Matsuda,N. and Mishina,M. (2004) A transposon-mediated gene trap approach identifies developmentally regulated genes in zebrafish. *Dev. Cell*, **7**, 133–144.